Research on Speech Intonation Recognition Technology Based on Deep Learning in Human-Computer Interaction

Jiajun Ge [1,*], Beize Lin [2], Jiaqi Zhang [3]

[1] Shanghai Guanghua Academy, Shanghai, China

[2] Ulster College, Shaanxi University of Science and Technology, Xi'an, China

[3] School of Computer Science, Xinyang Normal University, Xinyang, China

* Corresponding Author Email: gailuser189@gmail.com

Abstract. As a critical subfield of speech signal processing, speech intonation recognition technology aims to interpret paralinguistic features (such as pitch, rhythm, and energy) beyond the textual content of an utterance. Its development provides the core driver for enhancing the naturalness and emotional intelligence of human-computer interaction. This study focuses on intonation recognition technology, a critical component of speech signal processing. Its development has progressed from rule-based to statistical models, and now to deep learning models, resulting in steadily improving recognition accuracy. Regarding feature extraction, the acquisition of speech signal characteristics such as pitch, duration, and volume provides the data foundation for recognition models. Recognition algorithms have evolved from early Hidden Markov Models (HMMs) and Support Vector Machines (SVMs) to current mainstream deep learning models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). This technology is widely applied in areas such as voice assistants, intelligent customer service, and speech synthesis. For instance, it enables emotion analysis in voice assistants to adjust service strategies and enhances naturalness in synthesized speech. Current research focuses on developing algorithms robust to noise and accent interference while integrating with cognitive science. Future breakthroughs leveraging deep learning are anticipated in model complexity and recognition accuracy. Furthermore, driven by Internet of things and 5G technologies, applications are expected to expand into smart homes, telemedicine, and other domains.

Keywords: Speech Intonation Recognition, Deep Learning, Human-Computer Interaction (HCI), Speech Emotion Recognition (SER).

1. Introduction

With the widespread adoption of intelligent interactive devices, speech recognition has become a core method of human-computer interaction (HCI). Speech signals not only convey textual content but also carry multidimensional information such as emotional state, intent, and speaker identity through paralinguistic features. These features are crucial for enhancing the naturalness and contextual adaptability of interactions. However, most current Automatic Speech Recognition (ASR) technologies convert speech directly into text, often discarding these critical features, leading to misinterpretations of user emotions and intent. Potential solutions to overcome these limitations include generating semantic output fused with emotional labels directly from speech to avoid information loss, or incorporating visual cues for multimodal fusion to further improve accuracy.

Traditional intonation recognition primarily relies on manually designed rules or specific acoustic parameters (e.g., average pitch, maximum pitch), yielding limited effectiveness and requiring expert knowledge. Deep learning models have become central to the field, capable of understanding user intonation and emotions through extensive training on large datasets.

The world's largest emotion benchmark, EMONET-VOICE, released by the German LAION team in collaboration with the Technical University of Munich, contains 4500 hours of multilingual speech and labels 40 fine-grained emotions. This drives significant improvements in AI recognition accuracy for high-arousal emotions (e.g., anger, excitement) [1].

Companies like Tianrun Rongtong deploy customer service systems that identify user emotions (e.g., agitation, satisfaction) in real-time through speech intonation. These systems automatically adjust responses and prioritize complaints, reducing the need for human intervention by 30%.

The purpose of this review is to provide a comprehensive overview of speech intonation recognition technology based on deep learning and its application in human-computer interaction. The remainder of the paper is organized as follows: Section 2 elaborates on the methodological evolution of intonation recognition, including advancements in model architecture, feature extraction, and adaptive learning. Section 3 discusses current challenges and future prospects of the technology. Finally, Section 4 concludes the paper by summarizing research findings and outlining potential development directions [2].

2. Evolution of Deep Learning-Driven Speech Intonation Recognition Technology

The advancement of speech intonation recognition technology is deeply dependent on the innovation and application of deep learning methods. This section outlines recent representative progress in key areas such as model architecture, feature engineering, adaptive learning, and interactive system design, aiming to address limitations of traditional methods in accuracy, robustness, and contextual understanding.

2.1. Optimization of Foundational Model Architectures & End-to-End Learning

To address the complexity and high cost of traditional ASR models (e.g., GMM-HMM, DNN-HMM), which rely on separate training of pronunciation dictionaries, acoustic models, and language models, researchers are actively exploring more efficient end-to-end (E2E) solutions. Dai and Yang proposed the ConvTCN-FLASH-Transducer model, a lightweight E2E architecture specifically designed for Chinese. This model innovatively combines Convolutional Neural Networks (CNNs) and FLASH attention modules: multi-scale convolutions first capture local audio features, followed by Temporal Convolutional Networks (TCNs) extracting inter-frame temporal dependencies, strengthening local information coherence. Experiments on the THCHS30 Chinese dataset showed a significant reduction in Character Error Rate (CER) to 4.2%. To overcome the limitations of single loss functions (e.g., high computational complexity and slow convergence of Transducer Loss, or CTC Loss ignoring label dependencies), a strategy of joint training with Transducer Loss and CTC Loss was adopted. This effectively combines their advantages—preserving target sequence dependencies while improving training efficiency, model convergence speed, and stability [3].

2.2. Feature Engineering & Robustness Enhancement

Effective selection of high-dimensional features and robustness across languages and scenarios are core challenges in Speech Emotion Recognition (SER). Arruti et al. focused on emotion recognition for resource-scarce languages (Spanish and Basque), proposing a Feature Subset Selection (FSS) method based on Estimation of Distribution Algorithms (EDA). This method overcomes the poor performance of traditional greedy search in high-dimensional spaces by using probabilistic models to guide the search for optimal feature subsets. Validated using a three-stage progressive strategy (32/91/123 features) on the RekEmozio bilingual database, the EDA-FSS method significantly improved recognition accuracy (Basque: 80.05%, Spanish: 74.82%), achieving over 30% gains compared to no feature selection and traditional methods. The study also identified core prosodic and voice quality features robust across languages and genders, providing a basis for speaker-independent robust SER systems [4]. Geng Jianing's research investigated the effectiveness of feature extraction based on Mel-Frequency Cepstral Coefficients (MFCCs) combined with Convolutional Neural Networks (CNNs) for emotion classification (eight emotional dimensions). Cross-validation showed high accuracy (94%-97%) for emotions like happiness, disgust, and calmness. However, the study also revealed common limitations: 1) Poor dataset generalization:

Reliance on acted datasets degrades performance in real-world scenarios; 2) Insufficient cross-dataset stability: Dataset differences (emotion expression, recording conditions, labeling standards) weaken generalization; 3) Weak noise robustness: Environmental noise interferes with feature extraction and recognition accuracy, necessitating front-end denoising, robust feature/model design, or noise adaptation strategies; 4) High computational cost: Complex deep learning models struggle to meet real-time requirements on resource-constrained devices (e.g., IoT), demanding lightweight solutions [5].

2.3. Adaptive Learning for Specific Scenarios

The universality of recognition systems is often limited by acoustic variations in specific user groups or scenarios. Wu Ziteng's research addressed the common problems of acoustic and phonemic variation in patients with Mild Cognitive Impairment (MCI) by improving the traditional GMM-HMM system. The researchers innovatively proposed an adaptive learning method using weighted fusion of Maximum Likelihood Linear Regression (MLLR) and Maximum A Posteriori (MAP). This method effectively leverages limited MCI patient adaptation data to significantly enhance the acoustic model's ability to capture their unique acoustic characteristics, solving the acoustic variation problem. Simultaneously, by establishing a multi-pronunciation dictionary, it effectively handled common phonemic variations in MCI patients. Combining the optimized MCI-adapted acoustic model with the multi-pronunciation dictionary, a dedicated MCI patient speech recognition system was built. Experiments demonstrated that this system achieved higher accuracy in recognizing MCI patient speech compared to DNN-HMM and GPNN models [6].

2.4. Modeling and Implementation of Emotional Interaction Systems

The ultimate goal of speech intonation recognition is to enable natural and emotional human-computer interaction. Wang Yu's team proposed a Twin Finite State Machine (TFSM) based framework for emotional dialogue management. The core innovation lies in simultaneously constructing User FSM and System FSM models, dynamically simulating and driving the human-computer emotional dialogue process through information exchange between the two state machines. Compared to traditional single FSM, TFSM offers two main advantages: 1) Explicit separation of user and system states, better reflecting real interaction; 2) The system goal extends beyond task completion (e.g., information query) to actively respond and adapt to the user's emotional state, enhancing interaction naturalness and satisfaction [7]. Wang Jin's research designed and implemented a specific English speech recognition HCI system based on an improved. 1D Convolutional Neural Network (1DCNN). The system comprises four core modules: speech recognition, speech dialogue, video display, and speech processing. Key highlights include: 1) An anti-noise speech processing module using wavelet transform for speech information acquisition, classification, and denoising; 2) High-precision recognition of multiple emotions (happiness, anger, sadness, joy, etc.) in user speech, achieving 100% success rate in single-speech input scenarios and maintaining 90%-93.33% success even in complex scenarios with 2-5 mixed speech segments, significantly improving user experience [8].

3. Limitations and Challenges

Despite significant progress in deep learning-based SER for emotion recognition and HCI, the technology still faces key limitations and challenges in practical applications, with explainability being particularly prominent.

Explainability Challenge: The success of deep learning in SER largely stems from complex network architectures and massive training data. However, the inherent "black box" nature of Deep Neural Networks (DNNs) makes their internal decision-making mechanisms difficult to interpret clearly. While models often achieve high accuracy, the lack of transparency in the decision process poses problems. When the system makes an incorrect emotion judgment, users and developers

struggle to trace the error source, severely damaging system trustworthiness and user acceptance. A deeper issue is the model's inability to elucidate the specific contribution of different acoustic features (e.g., pitch, duration, energy changes) or contextual factors to the final emotion judgment, hindering targeted model tuning. Although eXplainable AI (XAI) methods like network activation visualization and feature importance analysis have been introduced and shown some progress, achieving the necessary depth of explanation for complex SER tasks requires further research.

4. Future Prospects

To address the aforementioned challenges, several technological pathways and application scenarios are envisioned. Speech recognition systems could integrate with large language models (e.g., ChatGPT) to generate real-time and emotionally appropriate responses. Combined with Text-to-Speech (TTS) technology that utilizes emotion tags to refine prosodic output, such systems could enable a new generation of voice assistants capable of serving as articulate and empathetic listeners [9].

In emerging applications, fine-grained emotional pulse charts—tracked at daily or weekly intervals—may help predict depression tendencies up to a week in advance, offering valuable tools for mental health monitoring. Similarly, in smart vehicle cockpits, systems may dynamically adjust in-cabin elements such as music, lighting, and haptic feedback through seat vibrations based on real-time emotion recognition, thereby mitigating driver frustration and reducing road rage incidents [10].

A proposed expert system, EMPATHIC-VOICE (Expert System for Paralinguistic Emotion Recognition in Human–Machine Emotional Interaction), exemplifies how hybrid approaches can merge interpretability with performance. Its core feature lies in a hybrid architecture combining a rule-based inference engine and a deep learning model (e.g., Whisper-Encoder with a lightweight MLP). The knowledge base incorporates an acoustic rule layer with over 350 hand-crafted IF-THEN rules (e.g., "IF Pitch Drop >12% AND Speech Rate >180 wpm THEN Emotion=Anger (0.8)"), a context layer accounting for user profiles, dialogue history, time, and noise, and an ethical rule layer enforcing GDPR compliance through data minimization, real-time anonymization, and automatic data erasure within 24 hours. The inference workflow processes real-time audio into 88-dimensional eGeMAPS features, predicts emotion probabilities via deep learning, and performs a second inference through the rule engine to output an emotion label with confidence and explanation. Target applications include intelligent customer service, smart cockpits, and companion robots for the elderly [11].

Furthermore, "Domain Adaptation + Transfer Learning" methods may serve as an emotion knowledge porter. A general speech emotion recognition (SER) model pre-trained on large-scale datasets (e.g., from customer service dialogues) can be adapted to specialized scenarios (e.g., in-car environments or nursing homes) using limited unlabeled or sparsely labeled data. Techniques such as feature alignment, adversarial training, and pseudo-label fine-tuning allow the model to efficiently learn domain-specific intonational patterns without requiring extensive new annotations [12].

5. Conclusion

Intonation recognition technology has made significant progress within the field of Natural Language Processing (NLP). From a technological evolution perspective, deep learning models have substantially enhanced recognition performance. At the application level, the technology is deeply integrated into numerous industries, tangibly improving human-computer interaction experiences and enhancing service efficiency and quality. However, current technology still faces challenges from noise, accents, and individual/cultural differences. Looking forward, efforts are needed on two fronts: 1) Continuously optimizing recognition algorithms to enhance robustness and generalizability; 2) Strengthening interdisciplinary integration, particularly delving deeper into human language generation mechanisms to further improve recognition accuracy. With the proliferation of

technologies like 5G and the Internet of Things (IoT), the application scenarios for intonation recognition will continue to expand. It holds promise for driving intelligent transformation in more fields, such as aiding communication for individuals with voice disorders, enabling smart home control, and facilitating remote medical diagnosis. This technology will propel speech interaction towards greater naturalness and intelligence.

Authors Contribution

All the authors contributed equally, and their names were listed in alphabetical order.

References

[1] Schuhmann C, Kaczmarczyk R, Rabby G, et al. EmoNet-Face: An expert-annotated benchmark for synthetic emotion recognition. 2025.

[2] Tianrun L, Xuan L, Guoxiang D. Intervention combined with surgical treatments for long-segment iliac artery occlusion. Chin J Minim Invasive Surg. 2025 Aug 24.

[3] Xin D, Yang S. End-to-end speech recognition based on ConvTCN–FLASH–Transducer.

[4] Álvarez A, Cearreta I, López JM, et al. Feature subset selection based on evolutionary algorithms for automatic emotion recognition in spoken Spanish and standard Basque language. Springer-Verlag; 2006.

[5] Geng J. Research on user-data-driven sentiment analysis.

[6] Wu Z. Design and implementation of an MCI human–machine interaction system integrating speech recognition and 3D emotional expression.

[7] Wang Y. Modeling methods and applications of emotion-aware dialogue management.

[8] Wang J. Design of an English speech recognition human–machine interaction system based on an improved 1D-CNN.

[9] Qin C, Zhang A, Zhang Z, et al. Is ChatGPT a general-purpose natural language processing task solver? arXiv. 2023; abs/2302.06476. doi:10.48550/arXiv.2302.06476.

[10] Li W. Research on the influence, recognition, and regulation of driver emotion and behavior in intelligent automotive cockpits [dissertation]. Chongqing: Chongqing University; 2021.

[11] Li Y, Li L. Research on the affective mechanism of harmonious human–computer interaction in network-based distance education systems. China Educ Info High Educ. 2015;(2):4.

[12] Mao L, Shi T, Wu L, et al. An unsupervised domain-adaptive text keyword extraction model—exemplified by texts in the “Artificial Intelligence Risks” domain. Inf Stud Theory Appl. 2022;45(3):6.